ABSTRACT

A method and a system for extracting information from a natural language text corpus based on a natural language query are disclosed. In the method the natural language text corpus is analyzed with respect to surface structure of word tokens and surface syntactic roles of constituents, and the analyzed natural language text corpus is then indexed and stored. Furthermore a natural language query is analyzed with respect to surface structure of word tokens and surface syntactic roles of constituents. From the analyzed natural language query one or more surface variants are then created, where these surface variants are equivalent to the natural language query with respect to lexical meaning of word tokens and surface syntactic roles of constituents. The surface variants are then compared with the indexed and stored analyzed natural language text corpus, and each portion of text comprising a string of word tokens that matches the any one of the surface variants or the natural language query is extracted from the indexed and stored analyzed natural language text corpus.

Elected for publication: Figure 1